“Identification of a novel drug target protein against Haemophilus influenzae Rd KW20: an insilico approach”

A Project

Thesis Submitted in Partial Fulfillment of

The Requirement for the Degree in

Bachelor of Technology

In Biotechnology Engineering

Submitted by:-

Om Bikash Kumar Das

111BT0571

Under the Supervision of:-

Dr. Nandini Sarkar

Assistant Professor

Department of Biotechnology and Medical Engineering

National Institute of Technology, Rourkela

Odisha-769008

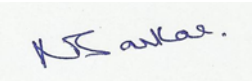
National Institute of Technology, Rourkela

Odisha-769008

CERTIFICATE

This is to certify that the project report entitle "Identification of a novel drug target protein against Haemophilus influenzae Rd KW20: an In silico approach" submitted by OM BIKASH KUMAR DAS (111BT0571) in the partial fulfillment of the required for the degree of the B.Tech in Biotechnology Engineering in Department of Biotechnology and Medical Engineering, National Institute of Technology, Rourkela is an authentic work carried out by him under my supervision. To the best of my knowledge the content in the report has not been submitted to any other Institute/University for any degree.

Date: - 8th May, 2015 Place: - Rourkela



Dr. Nandini Sarkar (Supervisor)

Assistant Professor

Department of Biotechnology and Medical Engineering

National Institute of Technology, Rourkela, Odisha-769008

ACKNOWLEDGEMENTS

I would like to convey the opportunity to extent my hearty gratitude to my guide and advisor Dr. Nandini Sarkar, Assistant Professor; Department of Biotechnology and Medical Engineering; National Institute of Technology-Rourkela,Odisha-769008, whose regular guidance and encouragement helped me a lot for the completion of my B.Tech thesis possible.

I would also like to thank my friend Ajeet Singh who helped me in my project work.

I would also like to thank National Institute of Technology Rourkela, for permitting to use all the required facilities in laboratories to carry out my project work.

Submitted by

Om Bikash Kumar Das

Roll No.-111BT0571

Department of Biotechnology & Medical Engineering

National Institute of Technology-Rourkela, Odisha-769008

Contents

List of Figures

List of Tables

Abstract:

Haemophilus Influenzae (H. Influenzae) is the gram negative bacteria causes infection at respiratory tract in human. Rd KW20 strain is mostly responsible for this disease. According to WHO statistics it kills 386,000 child per year in all over the world. In this approach we have identified some drug target protein which can be used as novel drug against this deadly pathogen. The metabolic pathways which are absent in the human but present in H. Influenza are taken as unique metabolic pathways. Here there are four such unique pathways are present only in case of bacteria, but not available in human. The genes present in these unique pathways were analyzed and listed on the basis of essentiality. These genes are crucial for survival of the pathogen and shortlisted from the Database of Essential Genes (DEG). The essential genes are blasted against the human genome through using BLASTP tool to shortlist the non-homologous genes. The gene named ponA, known as penicillin-binding protein is the best gene used for target against pathogen. The three-dimensional structure of this protein is predicted using Modeler 9.14, DeepView, RasWin and PyMol software. The active site for this gene is identified using CastP and the energically stabilized structure is chosen using Ramachandran plot.

Key word: Haemophilus Influenzae Rd KW20, ponA gene, 3D structure using Modeller 9.14, novel drug target,

1. Introduction:

Haemophilus influenza is a hazardous bacterial pathogen, bringing about respiratory tract infections in both kids and grown-ups [1]. This pathogen is present in nasopharynx, the upper respiratory tract of human body. It causes serious incursive infections to human body by extending the pathogen from nasopharnyx to the lower respiratory system. According to the survey done by World Health Organization (WHO), around 386,000 child deaths occur annually caused by H. influenza all over the world [2].

Haemophilus influenzae (H. influenzae) is a Gram-negative bacteria categorized to Pasteurellaceae family. It was first discovered in 1892 by Richard Pfeiffer. It is the first free living pathogen, whose entire genome project is sequenced and finished during 1995. It has both capsulated and unencapsulated strains. There are about eight different phenotypic characteristics and six different capsular antigen types, a-f categorized. Our current research is on Haemophilus influenzae Rd KW20 and to develop effective drug target using computational tools and technique. H. influenzae strain Rd KW20 has conventional been considered avirulent, when it cannot survive in the bloodstream of animals. The pathogen can be killed by normal adult human sera and very difficult to colonize the nasopharynx of infant rats. H. influenza strain KW20 is grown as monolayers of differentiated epithelium at the air liquid interface [3] & [4].

Several research work are going on progressively to develop the effective drugs by genetic or genomic approaches. Novel drug target are design to defend against antibiotic sensitive bacteria. New effective method has been developed in bioinformatics for finding organized targets antecedently from unexplored cellular functions and to empathize the inner biological process of pathogen. The complete genome information is also crucial for selection of accurate approach to check essentiality and selectivity pattern of the microbe. The target of the approach should be substantive encoded gene for the replication, growth and survival of pathogen. This target should not create any cytotoxicity damage to host. The genes called

as “essential genes” that are present in different conserved domain of genome and essential for the survival of the organism. These essential genes cannot endure inactivation through the mutation process [5]. The contingent pernicious mutants assist to adapt the status of these genes. Terminating the function of essential genes results death-dealing constitution inside bacteria. So it will be not worthless by addressing these drugs as “super bullet” against pathogen. This will not only help to avert cost but also very easy to detect virulent inhibitors by recognizing extend drug targets [2] & [5].

Now a days it’s very easy to recognize the targets by insilico-genomic approaches. “Differential genome method” is one of the beneficial approach for the anticipation of likely drug targets. This method offers detail genomic information of pathogens i.e. how the complete set of genes and protein are encoded inside the small genome [6] & [7]. The genes which present in pathogen, but absent in human are called non-homologous genes. These are most fundamental components for insilico-genome analysis. Using bioinformatics tools and techniques, the drug targets can be recognized so easily from these genes. The genes which are responsible for the foundation of life are known as the essential genes [8]. The function of essential genes are common to all cells. For the sustainment of infections is based to work out for anti-microbial agents against bacteria. The characterization of particular essential genes for specific pathogen can be used as drug target in several conserve domain of that bacteria. Database of Essential Genes (DEG) incorporates the list of essential genes of some limited pathogen. It is very easy to encounter the essentiality of genes after the successful development and implementation of human genome project databases. So it is tending to one step ahead development for novel drug target approaches. Anti-bacterial drug targets can be done by recognizing the specific essential genes by “subtractive genome approaches” [9], [10] & [11].

Subtractive genome approaches is successful implemented in this research paper to identify the potential drug targets for Haemophilus influenza. The essential genes for Haemophilus influenzae Rd KW20 are listed successfully by assisting Database of Essential Genes (DEG) against human genome. The genes present in Haemophilus influenzae Rd KW20, closely related to human genome are called as homologous genes and these genes are discarded [12], [13], [14] & [16].

The potential drug targets are effectively used in vaccination purposes. Vaccine provides procure immunity for the prevention of specific infection. Vaccine contains agents, which are part of an organism used to kill that organism. Vaccines may be toxins, surface proteins or inactive part of the organism which triggers the immune system to demolish the pathogen by identifying and recording the threat [17]. Operative vaccines can be developed by targeting the genes present in cell wall or plasma membrane.

Kyoto Encyclopedia of Genes and Genomes (KEGG) database provides unique metabolic pathway map of Haemophilus influenzae Rd KW20. As we are targeting the genes located in cell wall or plasma membrane, four important metabolic pathways are selected like c5-Branched dibasic acid metabolism pathways in Carbohydrate metabolism, Methane metabolism pathways in energy metabolism, Lipopolysaccharide biosynthesis and Peptidoglycan biosynthesis [18] & [19].

For novel antibiotic development ponA protein, which is also known as penicillin-binding protein of Haemophilus influenzae Rd KW20 is select for drug target. The structure of ponA can be predicted using various computational and bioinformatics approaches. Using Homology modeling, we can develop energycally stable three-dimensional structure for ponA protein [20], [21], [22], [23] & [24].

Protein achieves functional conformation by interacting with different molecules like ligand, substrate, DNA and other proteins. It's very crucial to obtain the specific three-dimensional protein structure for the

identification of proper interaction by visualizing the shape, physical, chemical and biological properties. By the enactment of protein surface characterization assist to analyze specification of binding, enzyme mechanism and examine for mutation.

Another important approach is by visualizing activity of protein using structure-based drug design (SBDD). The substrate binding site of protein helps in conformational changes and chemical modifications. This specific binding site of protein assist to trigger implementing the therapeutics approach for disruption in biological processes of pathogen.

1.1. Literature and review:

1.1.1 Life Cycle of H. Influanzae:

Interesting features about the cell structure of H. Influnzae; how it picks up energy; what essential molecules it it delivers. haemophilus influenzae is a microorganisms and consequently shows characteristics of a prokaryotic cell. It was distinguished as a gram negative microorganisms on account of its reaction to Gram staining techniques, as it stains red [1]. The gram negative coccobacillus has imperative cell wall components that assume a part in its survival and its pathogenicity. H. influenzae microbes comprise of different strains taking into account the presence or absence of an external covering called capsules. Haemophilus influenzae, the significant pathogen, can be differentiated into epitomized or typable strains, of which there are seven sorts (a-f) in light of the antigenic structure of the capsular polysaccharide, and unencapsulated or nontypable strains [2]. By segregating H. influenzae it was observed that some were indicated to have pili structures, which help in connection to the oropharyngeal

epithelial cell of human. Another essential properties of the H. influenzae cell structure is the rough lipopolysaccharide (LPS) which stretches out from the cell surface. There are varieties in the LPS from specie to specie and it has been recommended to be vital in the life cycle of the Haemophilus influenzae.

Haemophilus influenzae metabolizes sugar as its wellspring of vitality, however there is minimal thought about this metabolic ability of the H. influenzae. It is a facultative anaerobe and along these lines makes ATP by high-impact breath when oxygen is present and is likewise capable for metabolizing its sugar source without oxygen by fermentation. it was discovered that more than 90% of H. influenzae separated, digests sugars, for example, maltose glucose, galactose and ribose by fermentation and the remaining percent ferment fructose, mannose, or glycerol [3], [11] & [26].

Haemophilus influenzae reproduces by asexual procedure called binary fission which is characteristic to microscopic organisms. At binary fission, the H. influenzae starts replication at the source of replication site. As the chromosome is reproduced, proteins help in the development of the chromosome to inverse shafts of the cell and the extension of the cell. Septum formation and invagination of the cell layer divides the chromosomes into two different cells that are fit for developing to the shape of the first parent cell [3].

1.1.2 Mode of infection and symptoms:

H. Influnzae mostly affects the children below five years age. Haemophilus influenzae bacterias, are spread individual to-individual by direct contact or through respiratory droplets like by sneezing and coughing. Normally the microorganisms stay in the nose and throat-creating no problem. In some cases the microorganisms can enter the blood and spread, creating genuine disease in the person. More often

than not, Haemophilus influenzae microorganisms are spread by individuals who have the microbes in their noses and throats yet who are not sick (asymptomatic). The incubation period (time between first symptoms and exposure) of Haemophilus influenzae infection is not sure, but rather could be as short as a couple of days [3].

Infrequently Haemophilus influenzae microorganisms spread to other individuals who have had close or extensive contact with a patient with Haemophilus influenzae infection. In specific cases, individuals in close contact with that patient should get anti-microbial to keep them from getting the infection [1].

As of late there has been expanding recognition that this bacterium has a part in chronic lower inflammation of respiratory tract. However the interaction between H. influenzae and the lung is still not very much characterized. A combination of bacterial pathogenic character and deficiency of host defense may allow this bacterium to build contamination in the lower respiratory tract bringing about inflammation and clinical infection [9]. The other diseases caused by pathogen:

1. Bacteremia.
2. Pneumonia.
3. Epiglottitis.
4. Sinusitis.
5. Infectious arthritis.
6. Infect the host by attaching to the host using Trimeric Autotransporter Adhesins.

1.1.3 Earlier Therapeutic Approach:

Successful vaccines for Haemophilus influenzae have been discovered since the mid-1990s, and is suggested for kids under five age and asplenic patients. The World Health Organization suggests a

precautionary vaccine, consolidating vaccines against diphtheria, tetanus, pertussis, hepatitis B and Hib. There is not yet adequate confirmation on how viable this preventive vaccine is in connection to the individual vaccine [25], [26], [27], [28] & [29].

The available vaccines are very expansive compare to tuberculosis, diphtheria, measles, polio tetanus, and pertussis. Subsequently, though 92% of 92% of the populations of developed nations was vaccinated at the starting of 2003, vaccination scope was 42% for developing nations, and 8% for least-developed nations. The disadvantages of these vaccines are:

i. very expansive.

ii. Unfavorable reactions.

iii. Vaccine recipients ~30%.

iv. Causes swelling, or pain at the injection site.

1.1.4. Tools used for study:

NCBI:

Sequence alignment tools are used for comparability of amino acid sequences and characterized query genes. Basic Local Alignment Search Tool (BLAST) used to compare and quick search of protein and nucleotide sequences from databases. BLAST provides both local and global search alignment algorithm facilities to find the similarities from conserved domains of sequences. BLAST provides much faster alignment process implementing Smith–Waterman algorithm. There are five different version of BLAST like BLASTn, BLASTp, BLASTx, tBLASTn, tBLASTx. BLASTn assists to compare nucleotide sequences nucleotide databases. BLASTp assists to compare amino acid sequences from protein databases. BLASTx is used to compare six entrapped transcription product of a nucleotide sequences vs protein sequences. tBLASTx is used to compare six entrapped translation nucleotide sequence vs 6 entrapped sequence of nucleotide from database. tBLASTn is used to compare six entrapped translation nucleotide sequence vs six protein sequences from database [31].

KEGG:

Kyoto Encyclopedia of Genes and Genomes is a set of database of biological pathways, diseases, drugs, chemical substances, utilized for identification of genomics, metagenomics and metabolomics. It is an aggregation of pathway maps fusing various substances including qualities, proteins, RNAs, substance mixes, glycans, and compound responses, and furthermore infection qualities and targets, which are secured as individual doorways in exchange databases of KEGG [32].

DEG:

Database of Essential Genes, is a database and give tools to investigate the essentiality of the genes. Essential genes are those genes of an organism entity that are thought to be discriminating for its survival of the organism. Essential genes in a bacterium constitute a minimal genome, forming an arrangement of functional modules, which assume key parts in the emerging field, synthetic biology [18].

UNIPORT:

It gives data of the gene about the function, sequence and location in the cell. UniPort Knowledgebase is a protein database partially curated by specialists, comprising of two segments: UniProtKB/Swiss-Prot (containing assessed, manually annotated entries) and UniProtKB/ TrEMBL (containing reviewed, automatically annotated entries) [21].

CPHmodels (Computerized neural-system based protein demonstrating server):

CPHmodels is a gathering of databases and what's more, routines created to anticipate protein structure. It performs expectation of protein structure utilizing Comparative Modeling. It doesn't acknowledge more than 900 amino acids in the data succession. The arrangements are kept classified and are erased in the wake of preparing. This system did not issue me fitting results. The error it showed was like the one showed by Swiss Model [17], [30] & [31].

Swiss model:

It is used for automated homology modelling. It has a first approach mode that aides performs Homology Modeling. The user needs to enter his/ her email id and information the protein arrangement in Fasta position. It permits the user to pick as far as possible for format choice. It can seek the pdb document from the pdb database with the user giving the name of the pdb record or the client can transfer his/ her own pdb document. The yield record is a pdb document that is come back to the user's email address. The outcome can be sent by Swiss Model to PHD Secondary structure forecast at Columbia University furthermore, Fold Recognition Server (3D-pssm) of the ICRF [15] & [18].

Geno3D:

It performs Comparative protein structure modeling by spatial limitations (separations and dihedral) fulfillment. Geno3D is most habitually utilized for Homology or Comparative protein structure Modeling. Geno3d acknowledges information like Fasta organize yet just the one letter code must be utilized. The outcome is gotten in the PDB file format that can be seen in any Molecular Modeling software.Geno3d offers numerous other highlights, it permits the user to choose PDB entrances as formats for Molecular Modeling after a 3 stage iterative PSI BLAST. It exhibits the yield for every layout, alongside the optional structure forecast, shows percent of assertion in auxiliary structure and repartition of data from format on inquiry succession. The final result is sent to the user's email address. It likewise informs the client when

its server starts the Homology Displaying. It has an alternative where the user can choose what number of models to create. The fundamental thought behind having more than one model created is that the client may have a superior adaptability and comprehension. It likewise gives back a superimposed PDB document which has the models superimposed on one another. This is one of the great focuses in Geno3d as it permits us to think about the different models created in one window [22]. All the outcomes acquired can be downloaded as an archive.tar.Z that can be opened in WinZip in windows and in UNIX or Linux stages. So the user does not need to spare results in site page impact or in an archive record. It likewise shows the Ramachandran plot in the outcome.

Ramachandran plot:

The Sasisekharan-Ramakrishnan-Ramachandran plot describes permitted main chain conformations. A Ramachandran plot is an approach to visualize dihedral angles φ against ψ of amino visualize dihedral angles. It demonstrates the possible conformations of φ and ψ plots for a polypeptide. Rotation is allowed around the N-Cα and Cα-C single bonds of all residues (with one special case: proline). The angles φ and ψ around these bonds, and the angle of rotation around the peptide bond, ω, characterize the conformation of a residue. The peptide bond itself has a tendency to be planar, with two permitted states: Trans, ω ≈ 180° (generally) and cis, ω ≈ 0° (once in a while, and by and large at a proline deposit). The sequence of φ, ψ and ω points of all residues in a protein defines the backbone conformation [9].

MODELLER:

Modeler is used for homology or relative modelling of protein in three-dimensional structures. It is assembled in FORTRAN. It will runs on python script file commands. Modeler is most frequently utilized for homology or near protein structure demonstrating. Modeler aides focus the spatial limitations from the formats. It creates various 3D models of the arrangement you submit fulfilling the layout limitations. Modeler naturally figure a full molecule model. Modeler models protein 3D structure keeping in the requirements of spatial limitations. The restrictions can be gotten from various distinctive sources [29].

DeepView:

Swiss-PdbViewer is an application that gives an easy to use interface permitting to break down a few proteins in the meantime. The proteins can be superimposed in request to derive basic arrangements and look at their dynamic destinations or some other important parts. Amino corrosive changes, H-bonds, angles and separations between particles are anything but difficult to get because of the natural realistic and menu interface. DeepView - Swiss-PdbViewer was developed by Nicolas Guex (GlaxoSmithKline R&D). Swiss-PdbViewer is hard connected to SWISS-MODEL, an automated homology modelling server created inside the Swiss Institute of Bioinformatics (SIB) in Basel [10] & [11].

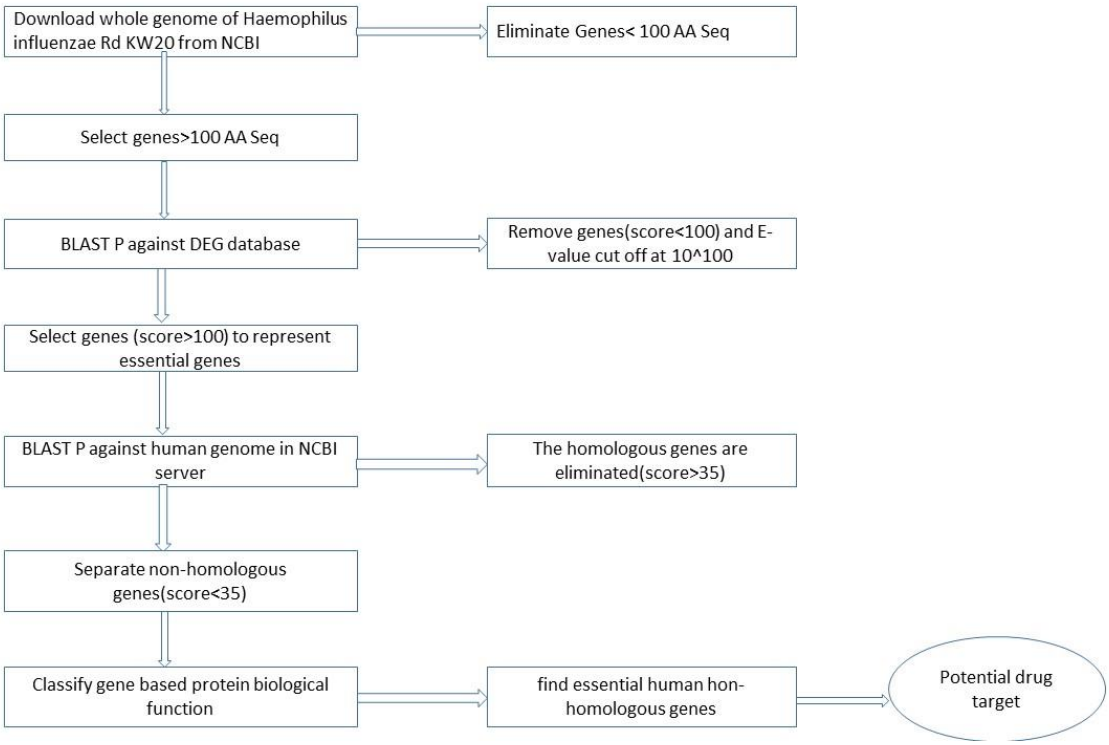
Chapter 2

OBJECTIVES AND WORK PLAN

2.1Objective

To identify a non-homologous essential gene of unique metabolism pathways, which can be used as potential drug target against Haemophilus Influenzae.

2.2Work plan



(Fig2The project plan in details)

Chapter 3

Material and method

3.1 Retrieval of proteome from NCBI:

The complete set of protein (proteome) is revived from NCBI. The sequence less than 100 amino acid sequence are consider to be paralog or duplicate protein. The non-paralog proteins are selected and paralog are eliminated.

3.2 Identification of unique metabolic pathways:

KEGG database is used for the selection of unique metabolic pathways from H. Influenzae and human. Some unique metabolic pathways are selected to identify appropriate genes. For the selection of genes in unique metabolic pathways, “pathway Entry” is selected. Then select “Metabolism” from drop down menu. The gene from the unique metabolic pathways are listed and analyzed.

3.3 Selection of essential genes:

To identify essential genes the amino acid sequence are submitted for BLASTP in **DEG (Database of Essential Genes)**. The above genes are analyzed through DEG database. To get the essential genes, cut off score greater than 100 are selected and non-essential genes are eliminated.

3.4 Identification of non-homologous genes:

Using BLASTP tool homologous and non-homologous genes can be differentiated. Homologous genes are present in both human and pathogen. Elimination of homologous genes are necessary, because these genes involves in the common biological processes and vaccination will be not effective. For the selection of essential non-homologous genes the identity is considered below 35 % and expected threshold value is set at 0.005. We are targeting the most conserved bacterial to get best result for multi resistant strain pathogen.

3.5 Homology modelling of identified protein:

The homologs conserved protein coding sequence was chosen from H. Influanzae strains for drug target. The three-dimensional structure of the targeted protein was displayed by considering the suitable all around contemplated protein structure is recognized by closeness search with the BLASTP tool against the protein databank. The homology modelling is done with online software like Geno3D , Swiss model, CPHmodels by using distinctive parameters. What's more, offline homology modelling is done utilizing profound parameters, the modeled protein was refined by the MODELER 9.14. The model is submitted

for the 3D-1D profile with VERIFY3D, and the stereo chemical qualities were checked with PROCHECK, Errat, Prove and WHAT_IF (http://nihserver.mbi.ucla.edu/SAVS/). At last, the basic properties of the target protein were visualized by using the Ramachandran plot score. The distinctive software models are contrasted and one another last best model is chosen; it is utilized for further drug design process.

3.6 Modelled protein structure validation using Ramachandran plot:

The best PDB result after the homology modelling is selected for Ramachandran plot analysis. The PDB file is submitted in SAVE (Structure Analysis and Verification) online server. The WHAT_CHECK tool of SAVE server will check the validation of protein structure. The result will be sand via web showing favoured, allowed and outlier region.

Chapter 4

Result and Description

4. Result and Description:

4.1 List of genes from unique metabolic pathways:

1. C5-Branched dibasic acid metabolism:

l

Gene Entry	Gene Details
HI0737	acetohydroxy acid synthase II; [EC:2.2.1.6]
HI1585	ilvI; acetolactate synthase 3 catalytic subunit [EC:2.2.1.6]
HI1584	ilvH; acetolactate synthase 3 regulatory subunit [EC:2.2.1.6]
HI1196	sucC; succinyl-CoA synthetase subunit beta [EC:6.2.1.5]
HI1197	sucD; succinyl-CoA synthetase subunit alpha [EC:6.2.1.5]
HI0988	leuC; isopropylmalate isomerase large subunit [EC:4.2.1.35 4.2.1.33]
HI0989	leuD; isopropylmalate isomerase small subunit [EC:4.2.1.35 4.2.1.33]
HI0987	leuB; 3-isopropylmalate dehydrogenase [EC:1.1.1.85]

2. Methane metabolism

Gene Entry	Gene Details
HI0185	adhC; alcohol dehydrogenase class III [EC:1.1.1.11.1.1.284]
HI0184	esterase; K01070 S-formylglutathione hydrolase [EC:3.1.2.12]
HI0007	fdxH; formate dehydrogenase subunit beta
HI0008	fdxI; formate dehydrogenase subunit gamma
HI0889	glyA; serine hydroxymethyltransferase [EC:2.1.2.1]
HI1556	glycerate dehydrogenase; K00018 glycerate dehydrogenase [EC:1.1.1.29]
HI0932	eno; phosphopyruvate hydratase [EC:4.2.1.11]
HI1636	ppc; phosphoenolpyruvate carboxylase [EC:4.1.1.31]
HI1210	mdh; malate dehydrogenase [EC:1.1.1.37]
HI0524	fba; fructose-bisphosphate aldolase [EC:4.1.2.13]
HI1645	fbp; fructose-1,6-bisphosphatase [EC:3.1.3.11]
HI0667	glpX; fructose 1,6-bisphosphatase II [EC:3.1.3.11]
HI0982	pfkA; 6-phosphofructokinase [EC:2.7.1.11]
HI1204	ackA; acetate kinase [EC:2.7.2.1]
HI1203	pta; phosphate acetyltransferase [EC:2.3.1.8]
HI0757	gpmA; phosphoglyceromutase [EC:5.4.2.11]
HI0465	serA; D-3-phosphoglycerate dehydrogenase [EC:1.1.1.95]
HI1167	serC; phosphoserine aminotransferase [EC:2.6.1.52]

HI1033	serB; phosphoserine phosphatase [EC:3.1.3.3]

3. Lipopolysaccharide biosynthesis

Gene Entry	Gene Details
HI1061	lpxA; UDP-N-acetylglucosamine acyltransferase [EC:2.3.1.129]
HI1144	lpxC; UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase [EC:3.5.1.108]
HI0915	lpxD; UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase [EC:2.3.1.191]
HI0735	UDP-2; K03269 UDP-2,3-diacylglucosamine hydrolase [EC:3.6.1.54]
HI1060	lpxB; lipid-A-disaccharide synthase [EC:2.4.1.182]
HI0059	lpxK; tetraacyldisaccharide 4'-kinase [EC:2.7.1.130]
HI1557	kdsA; 2-dehydro-3-deoxyphosphooctonate aldolase [EC:2.5.1.55]
HI1679	yrbI; phosphatase [EC:3.1.3.45]
HI0058	kdsB; 3-deoxy-manno-octulosonate cytidylyltransferase [EC:2.7.7.38]

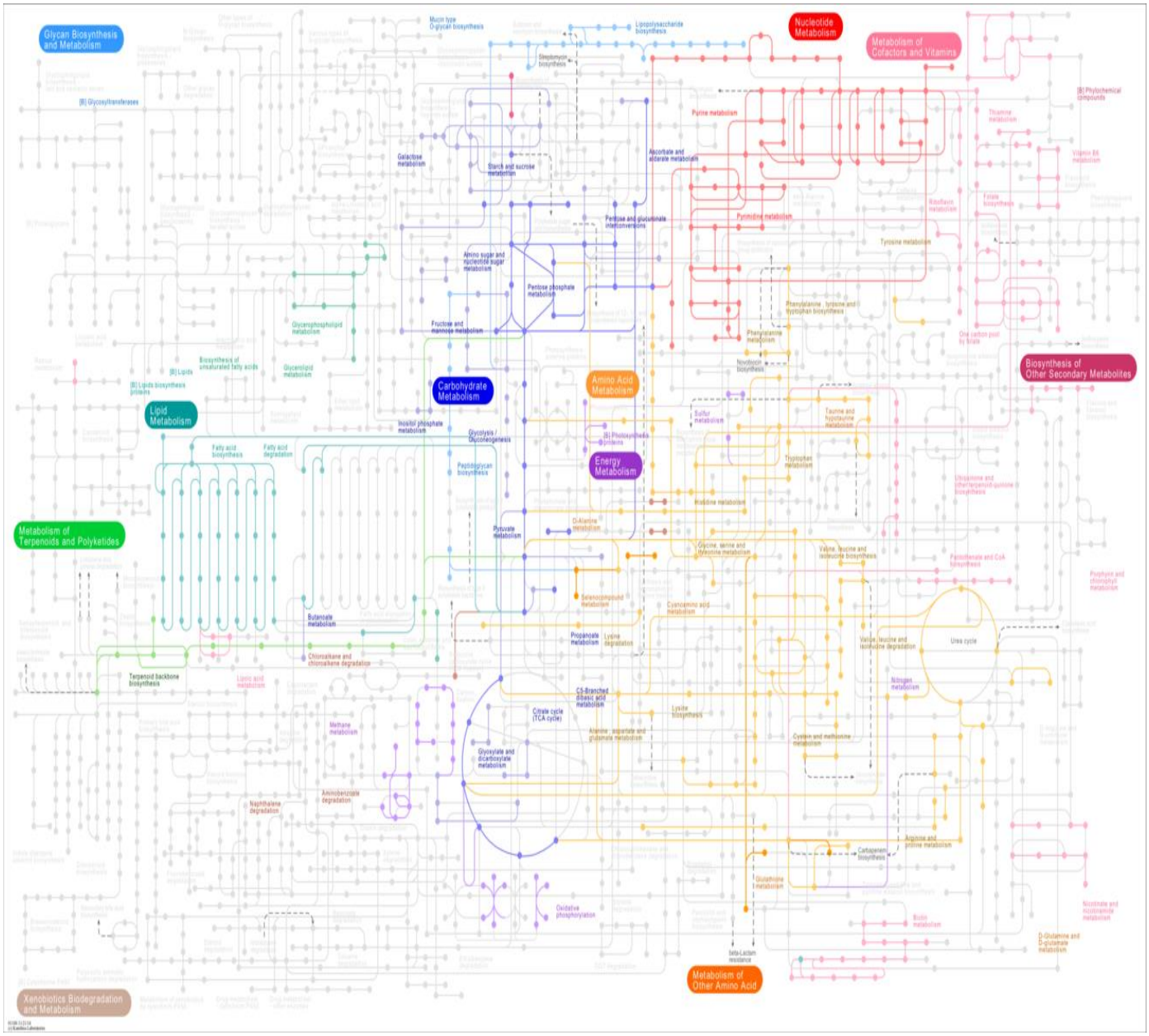
HI0652	kdtA; 3-deoxy-D-manno-octulosonic-acid transferase [EC:2.4.99.15 2.4.99.14 2.4.99.13 2.4.99.12]
HI1527	htrB; lipid A biosynthesis lauroyl acyltransferase [EC:2.3.1.-]
HI0199	msbB; lipid A biosynthesis (KDO)2-(lauroyl)-lipid IVA acyltransferase [EC:2.3.1.-]
HI0260.1	3-deoxy-D-manno-octulosonic-acid kinase; K11211 3-deoxy-D-manno-octulosonic acid kinase [EC:2.7.1.166]
HI1181	gmhA; phosphoheptose isomerase [EC:5.3.1.28]
HI1657	hypothetical protein; K03271 D-sedoheptulose 7-phosphate isomerase [EC:5.3.1.28]
HI1526	rfaE; bifunctional heptose 7-phosphate kinase/heptose 1-phosphate adenyltransferase [EC:2.7.7.702.7.1.167]
HI0621.1	D; K03273 D-glycero-D-manno-heptose 1,7-bisphosphate phosphatase [EC:3.1.3.833.1.3.82]
HI1114	rfaD; ADP-L-glycero-D-mannoheptose-6-epimerase [EC:5.1.3.20]
HI1105	rfaF; ADP-heptose-LPS heptosyltransferase II [EC:2.4.-.-]
HI0874	hypothetical protein; K02847 O-antigen ligase [EC:2.4.1.-]

4. Peptidoglycan biosynthesis

Gene Entry	Gene Details
HI1081	murZ; UDP-N-acetylglucosamine 1-carboxyvinyltransferase [EC:2.5.1.7]
HI0268	murB; UDP-N-acetylenolpyruvoylglucosamine reductase [EC:1.3.1.98]
HI1139	murC; UDP-N-acetylmuramate--L-alanine ligase [EC:6.3.2.8]
HI1136	murD; UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase [EC:6.3.2.9]
HI1133	murE; UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase [EC:6.3.2.13]
HI1140	ddl; D-alanine--D-alanine ligase [EC:6.3.2.4]
HI1134	murF; UDP-MurNAc-pentapeptide synthetase [EC:6.3.2.10]
HI1135	mraY; phospho-N-acetylmuramoyl-pentapeptide-transferase [EC:2.7.8.13]
HI1138	murG; undecaprenyldiphospho-muramoylpentapeptide beta-N-acetylglucosaminyltransferase [EC:2.4.1.227]
HI0964	mviN; virulence factor
HI0831	mtgA; monofunctional biosynthetic peptidoglycan transglycosylase [EC:2.4.1.-]
HI0440	ponA; penicillin-binding protein 1A [EC:3.4.-.- 2.4.1.-]

HI1725	ponB; penicillin-binding protein 1B [EC:3.4.-.-2.4.1.129]
HI0032	pbp2; penicillin-binding protein 2
HI1132	ftsI; penicillin-binding protein 3
HI0029	dacA; penicillin-binding protein 5 [EC:3.4.16.4]
HI1330	dacB; D-alanyl-D-alanine carboxypeptidase/endopeptidase [EC:3.4.21.- 3.4.16.4]

4.2 Identification of unique metabolic pathways:



(Fig.1. Reference pathway of H. Influnzae)

Identified unique pathways are listed below:

1. Carbohydrate metabolism

(a) C5-Branched dibasic acid metabolism pathway map.

2. Energy metabolism

(a) Carbon fixation pathway map in prokaryotes.

(b) Methane metabolism pathway map.

3. Lipopolysaccharide biosynthesis pathway map.

4. Peptidoglycan biosynthesis pathway map.

These five unique metabolic pathways are most important for vaccination purposes. Carbon fixation pathway map in prokaryotes is excluded because we are targeting the genes present in the location of cell wall or plasma membrane.

The unique pathway of KEGG website are as follows:

1. C5-Branched dibasic acid metabolism.

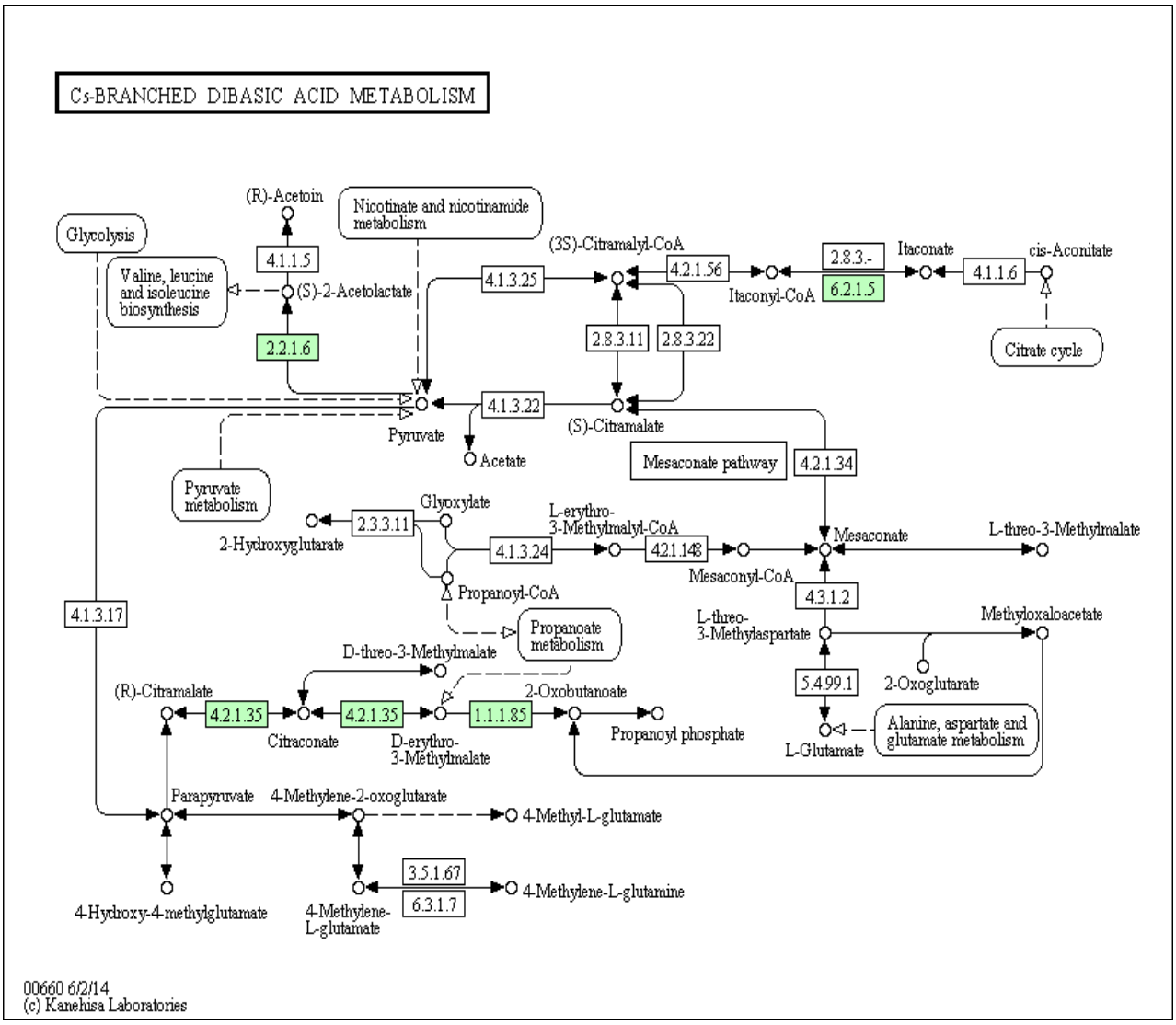
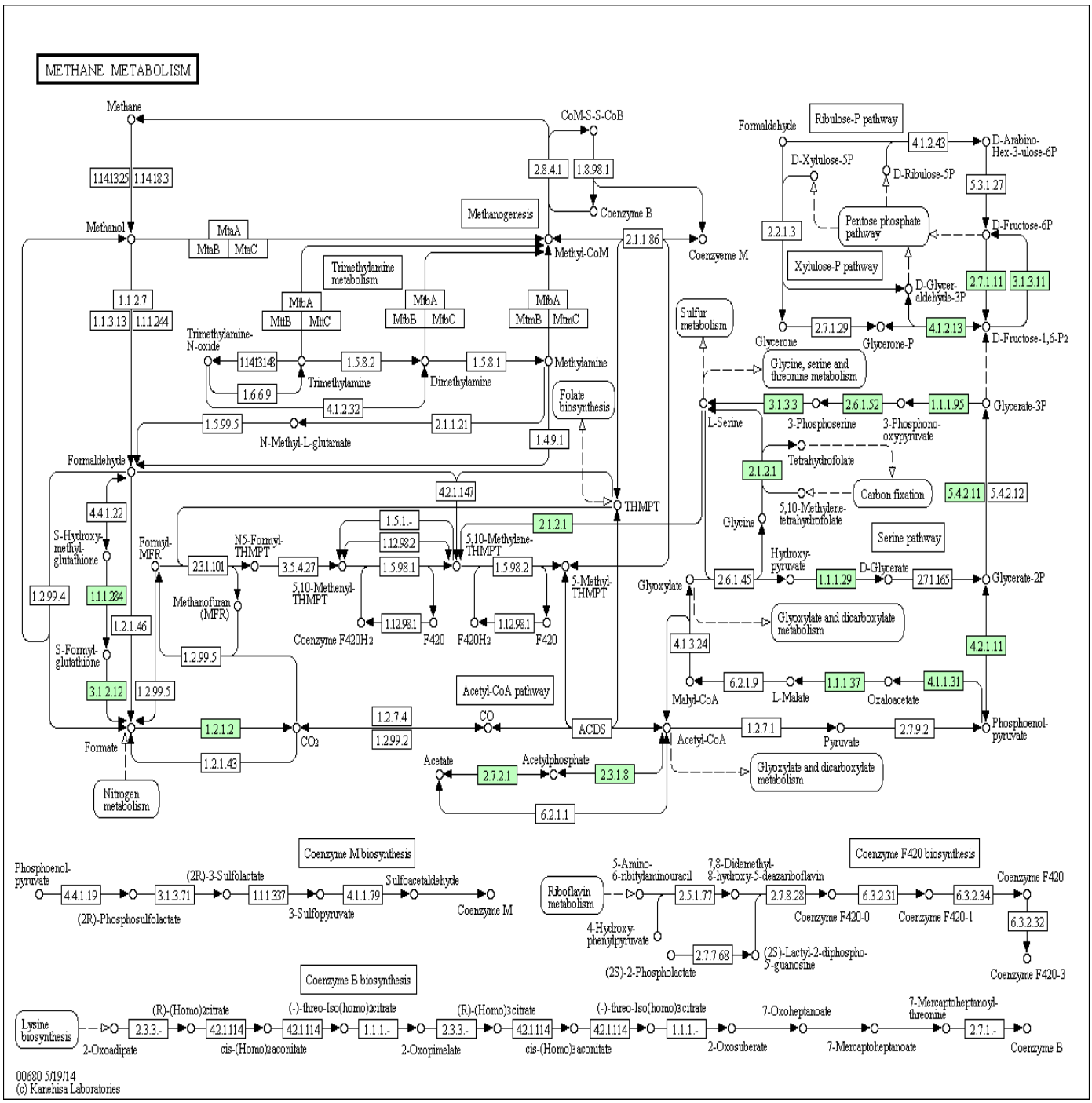


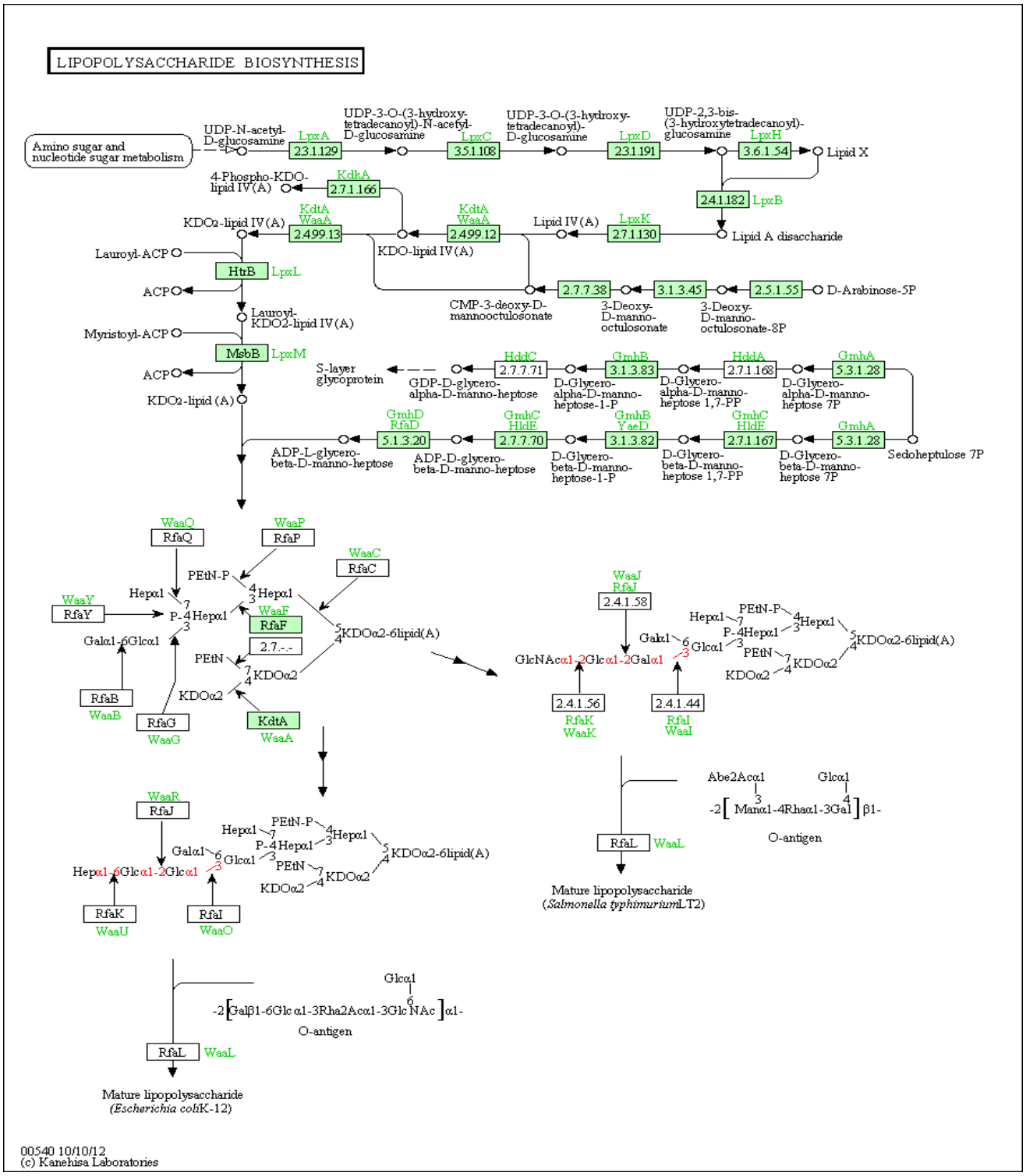
Fig.2. Carbohydrate metabolism

2. **Methane metabolism**:



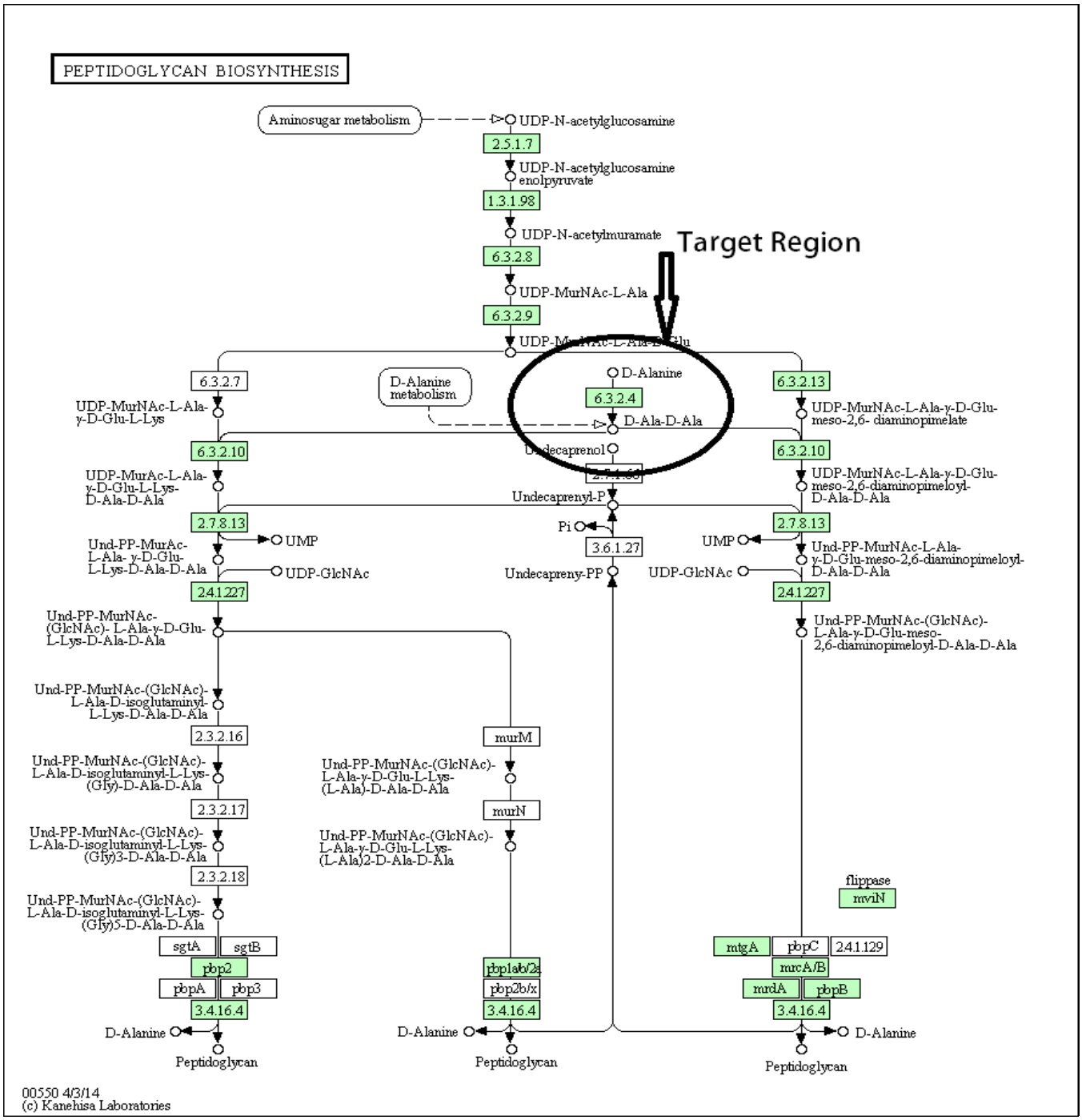
(Fig.3.Energy metabolism)

3. Lipopolysaccharide biosynthesis:



(Fig.4.Glycan biosynthesis and metabolism)

4. Peptidoglycan biosynthesis:



(Fig.5. Glycan biosynthesis and metabolism)

4.3 List of Essential Genes:

The essential genes are short out through the DEG database. The parameters that is being utilized to short out the genes is demonstrated as follows, score greater than 100(i.e. 500), expected value is $1 \times e0.5$. This is taken because of specificity of the gene toward the metabolism process.

After submitting all the amino acid sequence of the genes of selected four pathways are retrieved from the DEG database or through mail. All the essential genes are shown below.

Gene Entry	**Gene Name**
1. HI0737	Not available
2. HI0988	leuC
3. HI1196	sucC
4. HI1197	sucD
5. HI0008	fdxI
6. HI0184	Not available
7. HI0524	Fba
8. HI0889	glyA

9. HI0932	Eno
10. HI1033	serB
11. HI1167	serC
12. HI1204	ackA
13. HI0260.1	Not available
14. HI0735	Not available
15. HI1060	lpxB
16. HI1114	rfaD
17. HI1144	lpxC
18. HI1181	gmhA
19. HI1526	rfaE
20. HI1557	kdsA
21. HI1657	Not available
22. HI0029	Not available
23. HI0268	murB
24. HI0440	ponA
25. HI0964	mviN

26. HI1081	murZ
27. HI1135	mraY
28. HI1584	ilvH
29. HI1585	ilvI
30. HI1167	serC
31. HI1204	ackA

4.4 List of essential non-homologous genes:

The shortlisted 31 essential genes of H. Influenzae are subjected to BLASTP against human genome. The threshold value is set at 0.005 and identity less than 35% are considered.

After pasting the amino acid sequence and setting the parameters shown above, then select on BLASTP. This page will demonstrate the identity and detail information under description column section. Detail information and the most profitable record format of the gene can be attained to by making the gene in the provided box then clicking on the download or GenPept, graphical view for Graphics. We can decrease the number of column by tapping on the setting symbol on top of right corner.

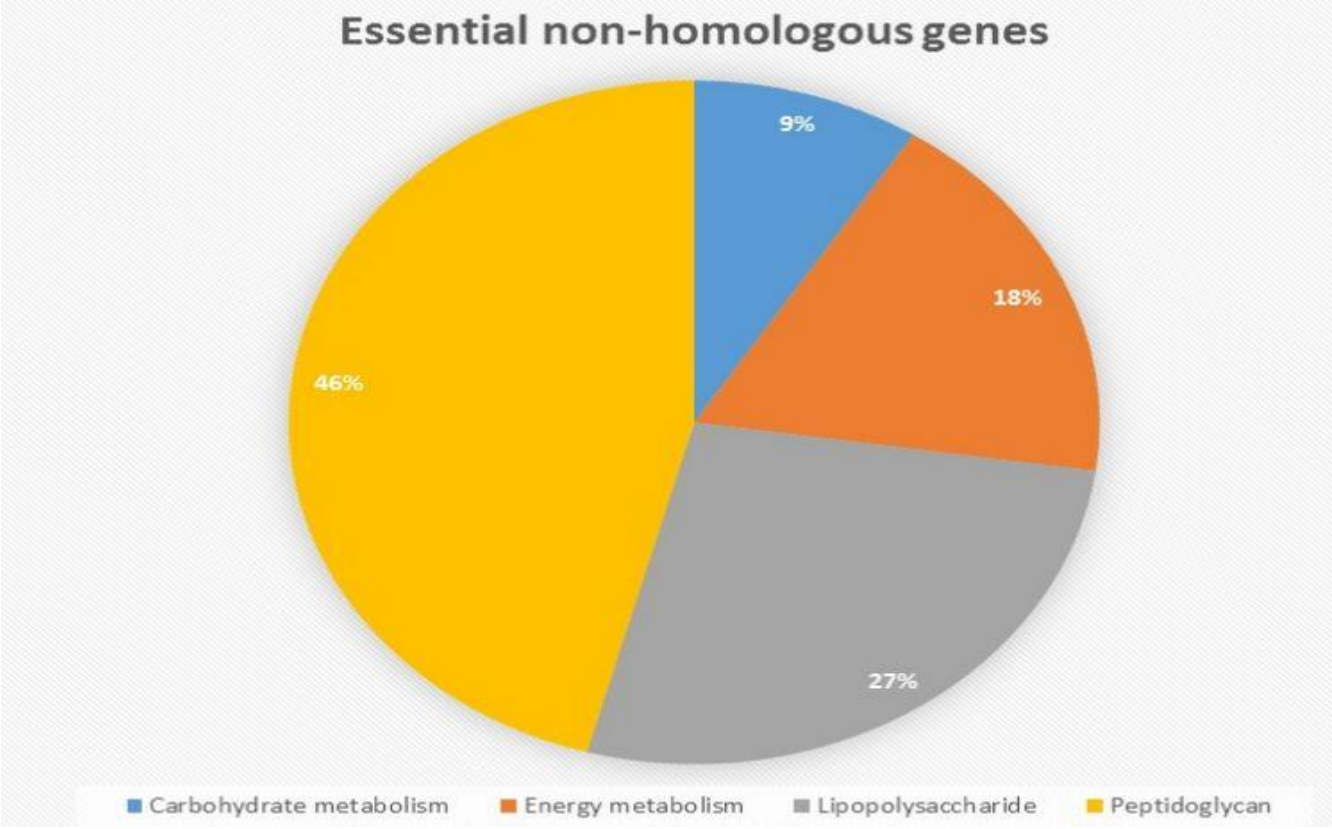
BLASTP results are listed below with their EC number and biological process:

	Accession No and Gene Name	Location in cell and cellular components	Can be used as drug or not	Biological Process	Enzyme Commission Number
1	HI0737	Cytoplasm	No	Truncated acetolactase synthesis	2.2.1.6
2	HI0988; leuC	-do-	Yes	Amino acid biosynthesis	4.2.1.33
3	HI1196; sucC	-do-	No	Tricarboxylic acid cycle	6.2.1.5
4	HI1197; sucD	-do-	No	Tricarboxylic acid cycle	6.2.1.5
5	HI0008; fdxI	Cytoplasmic Membrane	No	respiratory electron transport chain	Not Known
6	HI0184	Unknown	No	formaldehyde catabolic process	3.1.2.12
7	HI0524; fba	Cytoplasm, Periplasm, Cytoplasmic Membrane	Yes	Glycolysis	4.1.2.13
8	HI0889; glyA	Cytoplasm	No	Amino-acid biosynthesis, One-carbon metabolism	2.1.2.1
9	HI0932; eno	Cytoplasm, Periplasm, CytoplasmicMembrane	No	Glycolysis	4.2.1.11
10	HI1033; serB	Cytoplasm, Periplasm, Cytoplasmic Membrane	Yes	Amino-acid biosynthesis, Serine biosynthesis	3.1.3.3

11	HI1167; serC	Cytoplasm, Periplasm, Extracellular, Outer Membrane, Cytoplasmic Membrane	Yes	Amino-acid biosynthesis, Pyridoxine biosynthesis, Serine biosynthesis	2.6.1.52
12	HI1204; ackA	Cytoplasm	No	acetyl-CoA biosynthetic process	2.7.2.1
13	HI0260.1	Cytoplasmic Membrane	Yes	Lipopolysaccharide biosynthesis	2.7.1.166
14	HI0735	Cytoplasm, Periplasm, Cytoplasmic Membrane	No	Lipid A biosynthesis, Lipid biosynthesis, Lipid metabolism	3.6.1.54
15	HI1060; lpxB	-do-	No	Lipid A biosynthesis, Lipid biosynthesis, Lipid metabolism	2.4.1.182
16	HI1114; rfaD	-do-	Yes	Carbohydrate metabolism	5.1.3.20
17	HI1144; lpxC	-do-	No	Lipid A biosynthesis, Lipid biosynthesis, Lipid metabolism	3.5.1.108
18	HI1181; gmhA	-do-	No	Carbohydrate metabolism	5.3.1.28
19	HI1526; rfaE	-do-	Yes	-do-	2.7.1.167 2.7.7.70
20	HI1557; kdsA	-do-	No	Lipopolysaccharide biosynthesis	2.5.1.55

21	HI1657	Cytoplasm, Periplasm, Extracellular, Outer Membrane, Cytoplasmic Membrane	Yes	carbohydrate metabolic process	5.3.1.28
22	HI0029	-do-	Yes	Cell cycle, division, size, cell wall degradation and Peptidoglycan synthesis	3.4.16.4
23	HI0268; murB	Cytoplasm, Periplasm, Cytoplasmic Membrane	No	-do-	1.1.1.158
24	HI0440; ponA	Cytoplasmic Membrane	Yes	Resistance to Antibiotic, cell cycle, division, size, cell wall degradation and Peptidoglycan synthesis	2.4.1.- 3.4.-.-
25	HI0964; mviN	CytoplasmicMembrane	No	Cell cycle, division, size, cell wall degradation and Peptidoglycan synthesis	Not Known
26	HI1081; murZ	Cytoplasm, Periplasm, Cytoplasmic Membrane	No	-do-	2.5.1.7
27	HI1135; mraY	Cytoplasmic Membrane	No	-do-	2.7.8.13
28	HI1584; ilvH	-do-	No	Biosynthesis of Branched-chain amino acid	2.2.1.6
29	HI1585; ilvI	-do-	Yes	Biosynthesis of amino acids	2.2.1.6

30	HI1167; serC	-do-	No	Biosynthesis of amino acids Pyridoxine and Serine	2.6.1.52
31	HI1204; ackA	-do-	Yes	Not Known	2.7.2.1



(Fig.5.Essential non homologous genes in unique metabolic pathways)

4.5 Result of Homology modeling:

Three-dimensional structures will help in the visualization of the binding sites and may prompt the design of novel drug. The 3D structure of ponA protein of the H. Influenzae was modeled with Deep View; CPHmodels; Geno3D; Swiss model; Modeler 9.14 was used for fine building the model and global energy minimization.

Sl no	protein	Procheck	Verify3D	Errat
1	Geno3Dmodel	47.8	87.3	95.7
2	Deep view model	92.3	92.4	84.3
3	Modeller model	93.7	92.6	80.6
4	CPHmodels	89.4	94.2	92
5	Swiss model	89.7	88.9	89.5

(Modelling efficiency scores)

The above table demonstrates the modeler indicating preferred results over deepview, Swiss model. Modeler is the one of best homology modelling software. Modeler results are explained in details bellow:

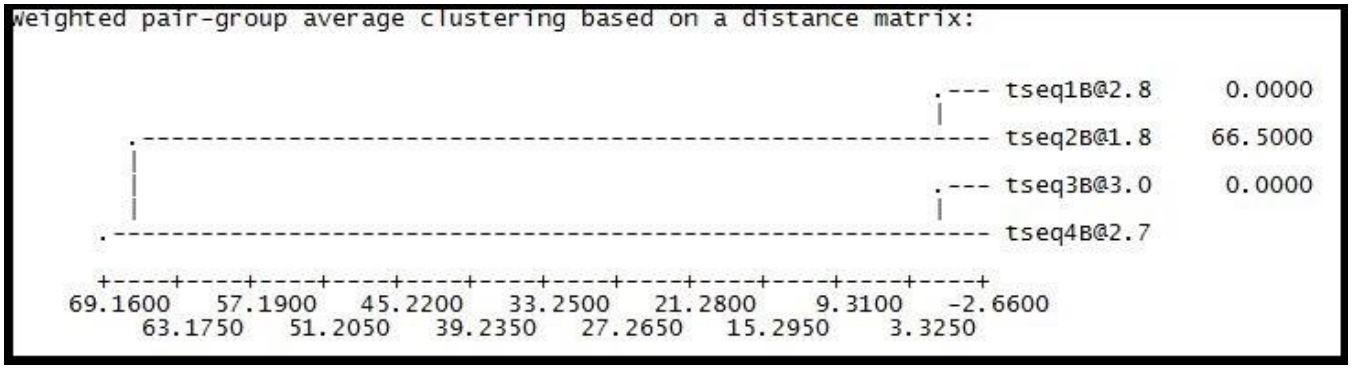


Fig 6 Pairwise distance matrix of clustering tree (dendrogram)

The first four best result from NCBI BLASTP results are named as tesq1, 2, 3&4. Tseq2 having pdb accession no 3UDF_A is showing best crystallographic structure. So it has higher crystallographic R-factor around 66.5 and sequence identity is around 41%.

There are five templet pdb files are generated and the best model is selected on the basic of DOPE score. The total number of residues of the model is 864 from 6723 number of selected real atoms. There are about 1192322 number of non-bonded pairs present in the model. The overall energy of the model is -49257.1602 Joule. Dope score are used to predict the most stable protein templet. Less is the DOPE score more is the stability and greater is the rank.

Rank of five protein templet are listed below on the basic of DOPE score:

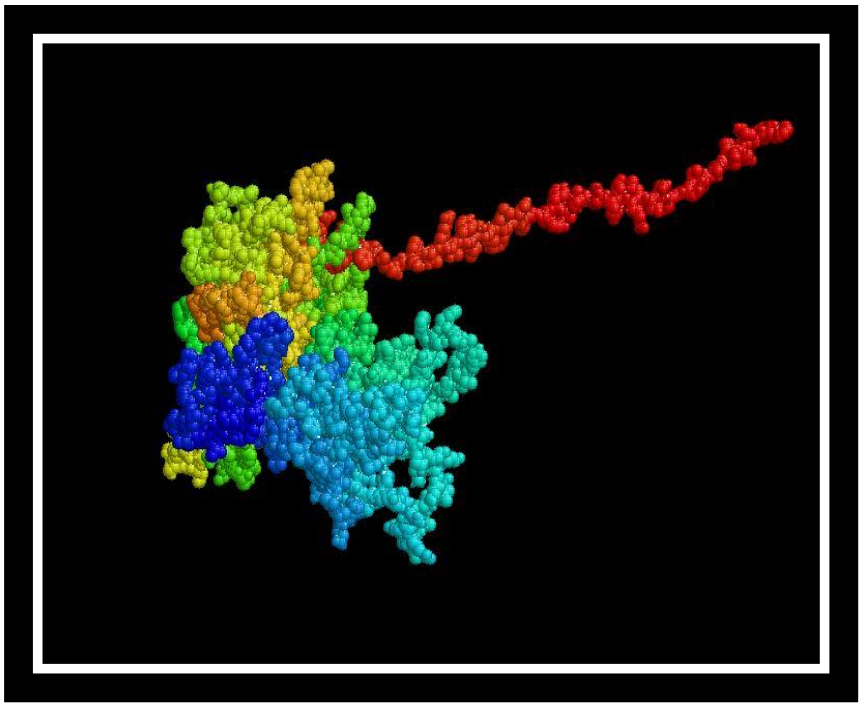
File name(pdb)	Identity	DOPE score	Rank
B99990001	**41%**	**-46309.45703**	**5**
B99990002	**40%**	**-50459.66406**	**2**
B99990003	**53%**	**-49490.78516**	**3**
B99990004	**53%**	**-49256.56250**	**4**
B99990005	**36%**	**-50748.95313**	**1**

List of energy parameters and values of protein from MODELLER:

Parameters	score
% sequence identity	33
Sequence length	864

Compactness	0.019779
Native energy (pair)	-1133.825974 J
Native energy (surface)	-188.171705 J
Native energy (combined)	-30.176544 J
Z score (pair)	-3.254623
Z score (surface)	-0.983342
Z score (combined)	-2.725979
Total DOPE score	-50748.953125 J

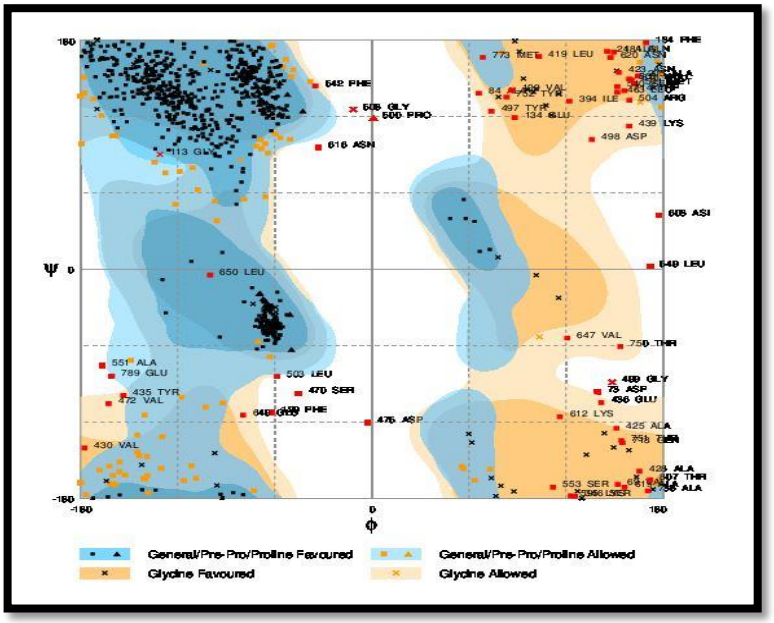
The three-dimensional surface structure of B99990005 is visualized using RasWin software. We can calculate the number of atom present in each side chain. Glutamic acid is present predominately in the protein. The position of each selected atom can be calculated using RasWin.

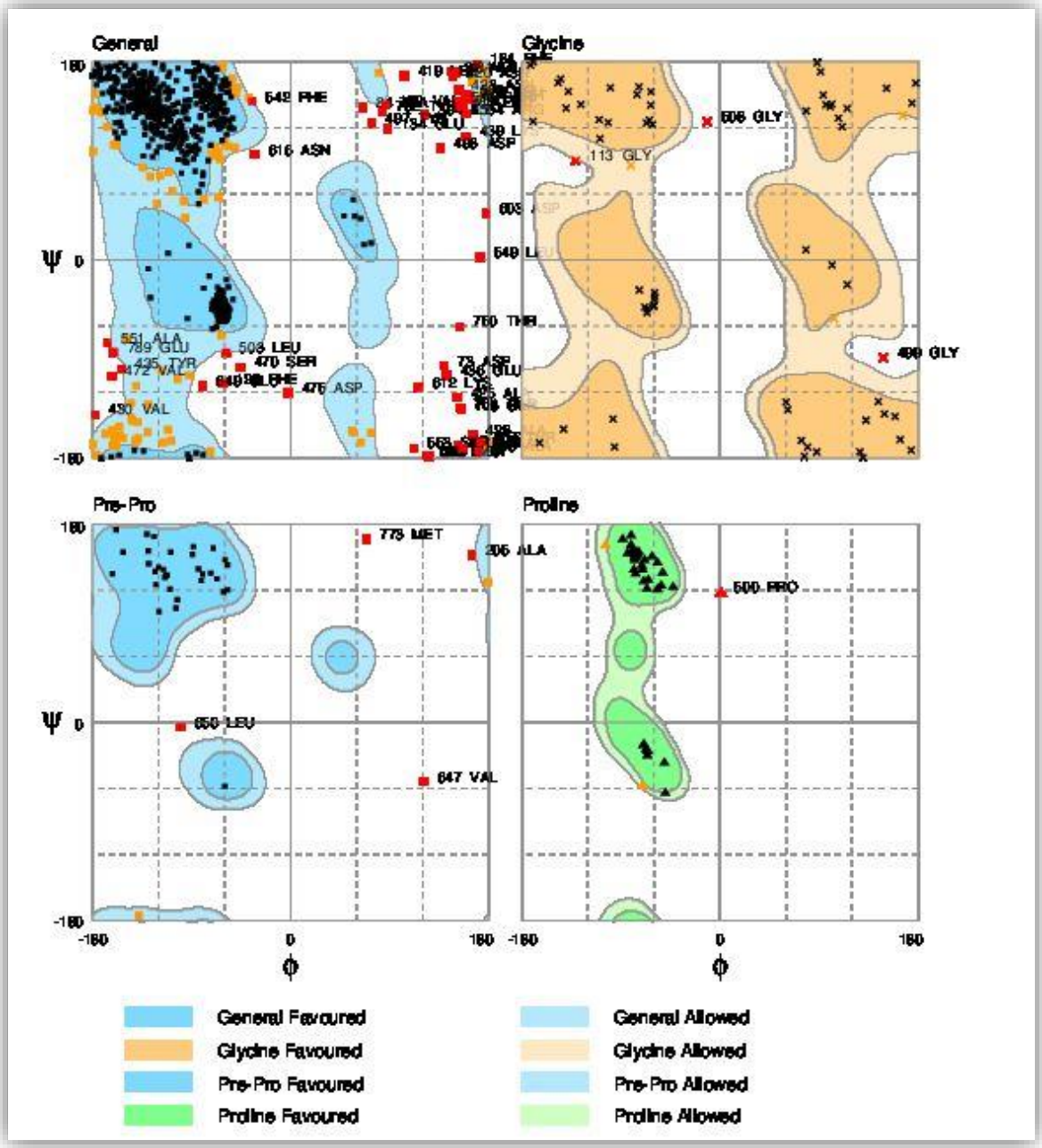


(Fig 7. 3-D Surface structure of B99990005 protein templet)

4.6 Ramachandran plot results:

The gene ponA shows stabilized structure and configuration of peptide bonds in the main chain N-Calpha and Calpha-C bonds.





(Fig 8.Ramachandran plot of ponA)

Residues found in favoured region: ~ 98 %(722 amino acids: 83.8%).

Residues found in allowed region: ~2 %(82 amino acids: 9.5%).

Residues found outlier region: 58 amino acids: 6.70%.

Chapter 5

Discussion

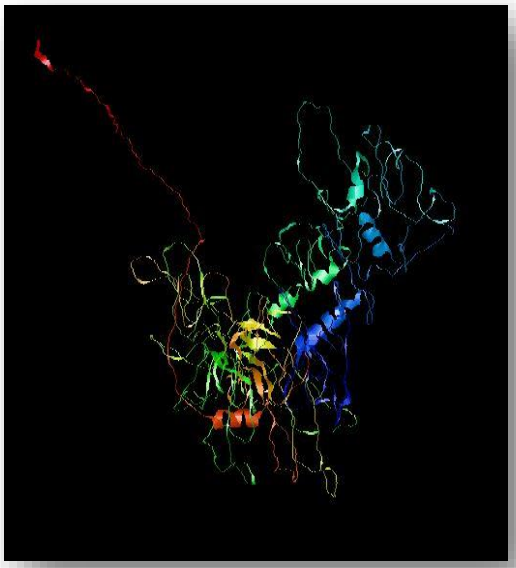
5. Discussion:

The ponA gene having NCBI Gene ID 949537 is identified as essential hon-homologs gene, is most preferable for vaccination purposes. This gene is also known as penicillin-binding protein of Haemophilus influenzae Rd KW20. This gene is present in the Cytoplasmic Membrane and involved in the biological process like resistance to Antibiotic, Cell size, Cell Lysis and Peptidoglycan synthesis. The chemical properties of the gene is similar to the modular pieces that form the peptidoglycan. When it is used as a drug target, blocks the enzymes that connect all the pieces together. The gene is constructed with long chains of sugars molecules with short peptides bonds sticking out in all directions. The D-alanyl-D-alanine carboxypeptidase region of the protein is cross-linked with these short peptides to form a three-dimensional structure. Acyl-ester intermediate is present in 441 position of the gene. It is the active region of the gene, because it helps in binding of metal ions. Metal ions like magnesium are crucial for drug targets approaches.

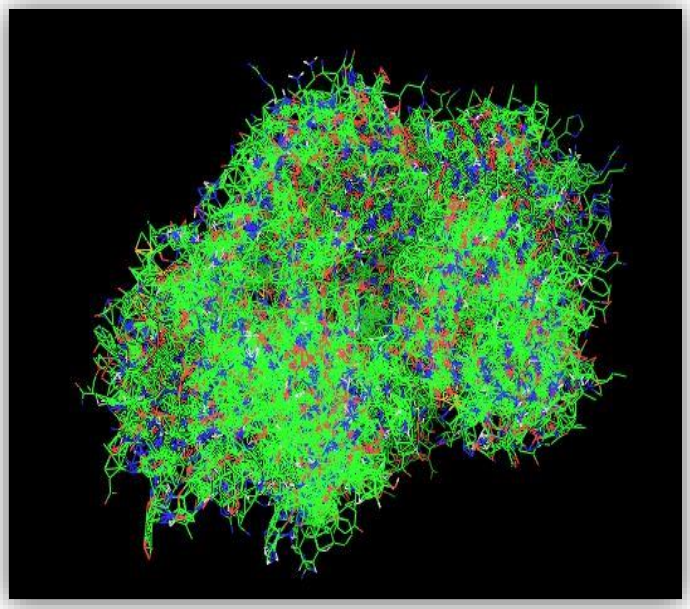
This can be taken as target protein, because of the following points:

1. The 3D structure of the protein is known.
2. It is the essential hon-homologous gene.
3. This gene is responsible for Peptidoglycan synthesis and Cell lysis process. So this gene will function effectively for potential drug target to disrupt the cell or plasma membrane.
4. It can block the metabolism process of pathogen, because it is not present in human.
5. The energy minimized structure is predicted.

Structure prediction of the gene ponA:



(Fig 9. The three-dimensnal structure of ponA, predicted using Pymol software)



(Fig 10.The energy minimized superimposed protein structure of ponA is calculated using DeepView software)

Chapter 6

Conclusion

6. Conclusion:

In this study, the genome of H. Influenzae from four important metabolism pathways were sucessfully analyzed, which are absent in the human. The essentiality of the genes were identified through the DEG tool. Around 31 genes are short listed from DEG. The essential genes were subjected for BLASTP against human genome. Using BLASTP homologous and non-homologous genes were separated. There was around 11 essential non-homologous genes, which can be used as drug target. After implementing all the steps successfully, we can able to identify a gene named as ponA for drug target. It is present in the Cytoplasmic Membrane of the pathogen. The pathogen H. Influanzae can be killed by blocking the biological function of ponA.

The future direction of this project is to perform Docking with ligands to the targeted protein, prediction of thermodynamic activities of ligands, and study about pharmacodynamics, pharmacokinetics, and solubility activities.

7. References:

1. A. E. Curr. (1998). New antibiotic discovery, novel screens, novel targets and impact of microbial genomics. Opin Microbiol. 1, 530-534.

2. Moxon ER: The molecular basis of pathogenicity in Haemophilus influenzae: comparative virulence of genetically-related capsular transformants and correlation with changes at the capsulation locus cap. Kroll JS, Microb Pathog 1989, 7:225-235.

3. Morens Fauci AS: Predominant role of bacterial pneumonia as a cause of death in pandemic influenza: implications for pandemic influenza preparedness. Taubenberger JK, J Infect Dis 2008, 198:962-970.

4. SCJ Lazaro E: Ampicillin-resistant non-beta-lactamase-producing Haemophilus influenzae in Spain: recent emergence of clonal isolates with increased resistance to cefotaxime and cefixime. Agents Antimicrob Chemother 2007, 51:2564-2573.

5. M. F.J Balzarini, J. Schools, (2007) "Broad Antiviral activity of Carbohydrate-binding agents against the four serotypes of dengue virus in monocyte-derived.

6. Sakharkar &Chow, V. T. K., (2004). A novel genomics approach for the identification of drug targets in pathogens, with special reference to Pseudomonas aeruginosa. In Silico Biol. 4, 0028.

7. Shaw, K.J & Hare, R.S., Vovis, G.F., (1999) Genomics and antimicrobial drug discovery. Antimicrob Agents Chemother. 43:439-446.

8. Huynen & Diaz-Lazcoz, Y. and Bork, P. (1997). Differential genome display. Trends Genet. 13, 389-390.

9. Ou, H.Y. and Zhang (2004). DEG: a Database of Essential Genes. Nucleic Acids Research. 32, D271-D272.

10. P. Rost, B. (1999). Twilight zone of protein sequence alignments. Protein Eng. 12, 85-94.

11. D. Ringe, (1995). What makes a binding site a binding site? Cur. Op. Struct. Biol.5,825.

12. Welch, W. and Jain, A.N. (1997). Automatic identification and representation of protein binding sites for molecular docking. Prot. Sci. 6,524.

13. M. L., (1983). Analytical molecular surface calculation. Journal of ApplCrystallogr. 16,548.

14. Marshall, G.R. (1990). De novo design of ligands. Journal Comput.- Aided Mol. Design, 4 337.

15. Stouten, P.F.W. (1995). Molecular Mechanics/Grid Method for the Evaluation of Ligand-Receptor Interactions. J. Comp.Chem. 16, 454-464.

16. R. S, (1999). Chlamydophila: intracellular biology, pathogenesis, and immunity. American Society for Microbiology. Washington, D.C.

17. Andersen (1998). Pathogenesis of lower respiratory tract infections due to Chlamydia, Mycoplasma, Legionella and viruses. Thorax. Apr. 53(4), 302-7.

18. Lesk (2005), Introduction to Bioinformatics, Second edition. Oxford University Press Inc., New York.

19. Ramachandran, V. (1963). Stereochemistry of polypeptide chain configurations. In: J. Mol. Biol. 7, 95-99.

20. Peitsch, M.C. (2003). SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res. 31, 3381-3385.

21. Lund and Brunak, S. (1997). Protein distance constraints predicted by neural networks and probability density functions. Protein Engg.10, 1241-1248.

22. Peitsch, M.C. (1999). Protein modelling for all. TiBS. 24,364-367.

23. G, Vriend (1990). WHAT IF: a molecular modeling and drug design program. J. Mol. Graph. 8, 52-56.

24. Bowie, J. U. (1997). VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol. 277, 396-404.

25. Stouten, G. P, Jr (2000). Fast Prediction and Visualization ofProtein Binding Pockets with PASS. Journal of Computer-Aided Molecular Design, 14, 383-401.

26. A. Singh, S. K., Ghosh, and Bandyopadhyay, (2006). In silico identification of potential therapeutic targets in the human pathogen Helicobacter pylori. Biol. 6, 0005.

27. O.N. (2006). ACD/ChemSketch Freeware, version 10.00, Adv. Chemistry Development, Inc., Canada.

28. Olson, A. J. (1998). Automated Docking Using a Lamarckian Genetic Algorithm and Empirical Binding Free Energy Function. J. Computational Chemistry. 19, 1639- 1662.

29. S.d. frank (1991). Control of protein phosphatase 2A by simian virus 40 smalltantigen.

Mol. Cell Biol. 11(4): 1988-1995.

30. C.P H, Shinagawa and Morikawa, K. (2001). Crystal structure of the Holliday junction migration motor protein RuvB from Thermus thermophilus HB8. Proc.Natl.Acad.Sci. 98, 1442-1447.

31. SD Jain C.D., Clancy, S.B, H., Gonzalez, S., Wetmur, J.G .and Tainer, J.A. (2001). Structure and mechanism of the RuvB Holliday junction branch migration motor. J. Molecular Biol. 311,297-310.

32. JP., Cann, I.K.O., Ishino, S.P. and Morikawa, K. (2001). Atomic Structure of the Clamp Loader Small Subunit from Pyrococcus furiosus. Mol. Cell. 8, 455-463.